analytical chemistry

Article

Cite This: Anal. Chem. 2017, 89, 11999-12005

pubs.acs.org/ac

Machine-Learning-Based Olfactometer: Prediction of Odor Perception from Physicochemical Features of Odorant Molecules

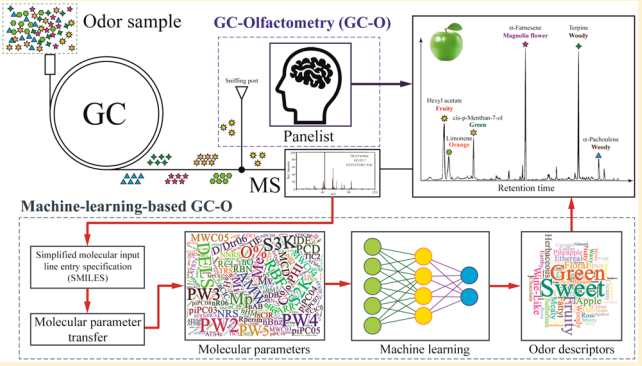
Liang Shang,[†] Chuanjun Liu,[†,§] Yoichi Tomiura,[‡] and Kenshi Hayashi*[,†]

[†]Department of Electronics, Graduate School of Information Science and Electrical Engineering, and [‡]Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan

[§]Research Laboratory, U.S.E. Company, Limited, Tokyo 150-0013, Japan

Supporting Information

ABSTRACT: Gas chromatography/olfactometry (GC/O) has been used in various fields as a valuable method to identify odor-active components from a complex mixture. Since human assessors are employed as detectors to obtain the olfactory perception of separated odorants, the GC/O technique is limited by its subjectivity, variability, and high cost of the trained panelists. Here, we present a proof-of-concept model by which odor information can be obtained by machine-learning-based prediction from molecular parameters (MPs) of odorant molecules. The odor prediction models were established using a database of flavors and fragrances including 1026 odorants and corresponding verbal odor descriptors (ODs). Physicochemical parameters of the odorant molecules were acquired by use of molecular calculation software (DRAGON). Ten representative ODs were selected to build the prediction models based on their high frequency of occurrence in the database. The features of the MPs were extracted via either unsupervised (principal component analysis) or supervised (Boruta, BR) approaches and then used as input to calibrate machine-learning models. Predictions were performed by various machine-learning approaches such as support vector machine (SVM), random forest, and extreme learning machine. All models were optimized via parameter tuning and their prediction accuracies were compared. A SVM model combined with feature extraction by BR-C (confirmed only) was found to afford the best results with an accuracy of 97.08%. Validation of the models was verified by using the GC/O data of an apple sample for comparison between the predicted and measured results. The prediction models can be used as an auxiliary tool in the existing GC/O by suggesting possible OD candidates to the panelists and thus helping to give more objective and correct judgment. In addition, a machine-based GC/O in which the panelist is no longer needed might be expected after further development of the proposed odor prediction technique.



Gas chromatography/olfactometry (GC/O) has been developed as a powerful tool in the field of odor research because of the coupled performance of gas chromatographic analysis with human panelist sensory detection.[1,2] GC/O can work not only as an instrumental analysis to identify and quantify complex odor mixtures but also as a sensory analysis to assess odor or odor-active compounds within the GC effluents.[3] In GC/O analysis, the eluted substances are perceived simultaneously by two detection systems; one is a mass spectrometry (MS) system and the other is the human olfactory system.[4,5] Evaluation by a human sniffer plays an important role because it can make up for deficiencies of GC (or GC/MS) in odor analysis.[6] For example, many of the peaks detected by GC for an odor mixture may not actually contribute to our perception since they are present below our thresholds for detecting them. Conversely, some compounds may not show up as detectable GC peaks but may have a low perception threshold and contribute substantially to a sample's profile. The sensory evaluation of smells by trained panelists can overcome such problems and represents a valid approach to odor assessment. Through sniffing GC effluent components, panelists can determine the odor characteristics ascribed to each individual component, which is important information for the overall odor analysis.[7]

A major problem of GC/O is the subjectivity of assessors at the intra- and inter-individual level. Sensory assessment of smells by panelists is influenced by many factors, such as the testing environment, experimental bias, assessor sensitivity, assessor selection, and training.[8] Experimental conditions should be well-established to ensure accuracy and precision of the odor descriptor data collected by the panelist. Therefore, although GC/O has presented many challenges not considered in typical GC analysis, its application and promotion are hindered by the variability, high technical requirements, and

Received: June 21, 2017
Accepted: October 12, 2017
Published: October 13, 2017

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005

high costs of the trained panelist.[9−11] Some software modules have been designed and applied in GC/O as a supplement to odor and chemical analysis. For example, AroChemBase (Alpha MOS) constitutes the most comprehensive chemical and sensory library ever, which is convenient for fast sensory profiling and detailed chemical/odor characterization. In this kind of software module, however, the number of compounds with odor descriptors (around 2000) is far less than that of the total compound pool (around 44 000). Therefore, regardless of human assessment or software indexing, the question of how to effectively obtain sensory information for eluted compounds from GC is still unanswered for GC/O.

Recently, there has been growing interest in the prediction of the structure−odor relationship by use of various parameters of odorant molecules, such as structural, topological, geometrical, electronic, and physicochemical.[12] The driving force of this research might come from the great progress made by Buck and co-workers[13] and Axel and co-workers[14] through their discoveries of olfactory receptors and the organization of the olfactory system. It has been revealed that odorants with similar features in molecular profile show similar activity patterns in the olfactory bulb (also referred to as the odor map).[15,16] Great efforts have been made in odor classification from different aspects such as semantic classifications, olfactory descriptors, odor similarities, statistical analysis, and olfactory profiles.[17−21] Although it is still difficult to establish a general rule to predict the olfactory perception of odorants, a number of computational techniques have been used successfully in the explanation of both physicochemical and perceptual spaces of odorant molecules. For example, Sobel and co-workers[22] related these two spaces to each other and find that the primary axis of perception (defined as odor pleasantness), reflects the primary axis of physicochemical features. Kumar et al.[23] developed a network-based approach (smell network) that can be used to explore the perceptual universe and prove the underlying similarity of percepts. Keller et al.[24] established a machine-learning algorithm using a large olfactory psychophysical data set, which can be used to predict odor intensity, pleasantness, and semantic descriptors from chemical features of odor molecules. These developments demonstrate that it becomes realistic to predict olfactory perception from structural parameters of odorant molecules. Additionally, the development of data analysis techniques and cheminformatics software make it possible to deal with the complexity of odor spaces that have high dimensionality and nonlinearity, which would also be applied in many fields such as food and fragrance evaluation.[25−27]

However, to our knowledge no study has applied the odor prediction models in GC/O. In this proof-of-concept study, we test the possibility that a machine-learning-based prediction model could be used to replace the human panelist in GC/O. As illustrated in Figure 1, after the GC effluent is identified by mass spectrometry, its molecular parameters (MPs) can be transferred by a cheminformatics software and inputted into a classifier system in which each classifier is labeled by a specific odor descriptor (OD) (a word like sweet, green, fruity, herbaceous, etc.). After the true or false classification, the system can output the sensory information on the GC effluent, which may consist of single OD (such as sweet) or multiple ODs (such as sweet and green). These ODs predicted by these models would be regarded as references for odor sensory information evolution.

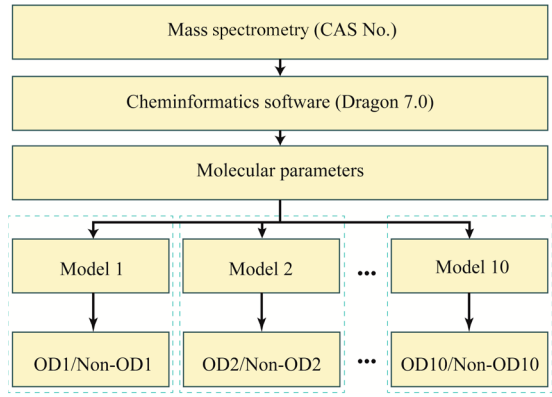


Figure 1. Concept diagram to predict odor descriptors by using molecular parameters.

A flavor and fragrance database (Sigma−Aldrich, 2016) that includes 1026 odorants and 10 ODs was considered in this study. The physicochemical MPs were acquired via cheminformatics software. The features of the MPs were extracted via either unsupervised (principal component analysis, PCA) or supervised (Boruta, BR) approaches. Ten typical ODs with high frequency of occurrence in the database were selected to establish the models. Different machine-learning algorithms, including support vector machine (SVM), random forest (RF), and extreme learning machine (ELM), were used and their prediction results were compared. Finally, Golden Delicious apple GC/MS data were employed to prove the feasibility of the model calibrated in present study. A Boruta-SVM model showed high accuracy in OD prediction, which indicates the possibility for machine-learning based GC/O.

MATERIALS AND METHODS

Odor Data Collection. Simplified molecular input line entry specification (SMILES) was obtained by both semi-automatic and manual methods from PubChem (https://pubchem.ncbi.nlm.nih.gov) according to the CAS number of the odorant molecules recorded in the Flavors and Fragrances database.[28] The SMILES strings were imported into the Dragon chemoinformation software (version 7.0, Kode, Italy) to compute the physicochemical parameters. The calculation afforded 5270 parameters with various values for each odorant molecule. It was found that most of parameters (around 4200) were assigned as not applicable (NA). We removed these parameters with NA and finally got a parameter matrix with 1006 MPs. All MPs were normalized and centered for further processing.

Data Analysis. The data analysis process is shown in Figure 2. The data set for odor prediction is a typical imbalanced data set because the class distribution of positive samples (minor samples with specified OD labels) and negative samples (major samples with nonspecified OD labels) is not uniform. Here, synthetic minority oversampling technique (SMOTE) was employed to overcome the imbalance problem.[29] The minority class was oversampled at 300% of its original size and the majority class was undersampled to obtain a balanced data set. Afterward, the sample pool was divided into training and test sets with a 3:1 ratio by use of the Kennard−Stone (KS) algorithm.[30,31] Sample size details for each OD are listed in Table 1. The unsupervised feature combination method (PCA) and supervised feature selection method (Boruta) were performed to extract kernel information to enhance the

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005

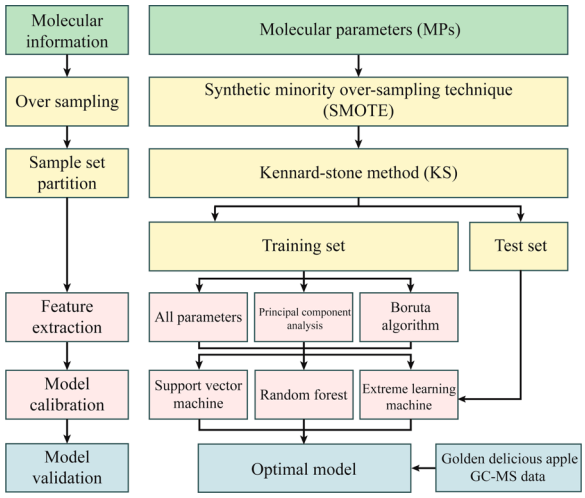


Figure 2. Data processing diagram of prediction model calibration and validation.

performance of the classification frameworks.[32,33] SVM, RF, and ELM classification algorithms were applied to predict ODs. The optimal model was determined by considering the accuracies of the training and test sets. As the last step, the F1 score based on precision and recall was used to verify the performance of the optimal model.[34]

Model Validation. For validation of the OD prediction models developed in the present study, volatile compounds identified from actual samples were analyzed were employed. Arvisenet et al.[35] studied primary volatile organic compounds (VOCs) from Golden Delicious apples by solid-phase microextraction (SPME) and GC/MS analysis. Their results indicated that 30 compounds, including 13 esters, nine alcohols, five aldehydes, one ketone, one phenol, and (*E*,*E*)-α-farnesene, would be considered as primary. Based on the MPs calculated by Dragon 7.0, the optimal model (BR-C-SVM) calibrated in the present study was applied for predicting their ODs. Compared with the ODs reported in other papers or databases, the OD prediction models developed in this study would be evaluated.

RESULTS AND DISCUSSION

Odor Descriptors. It is well-known that, for machine learning, the larger the sample size is, the higher the model accuracy tends to be. For odor prediction, an optimum database should have an appropriate number of odorant molecules and ODs. Up to now, there are a number of odor databases that have been reported and analyzed.[36] Recently, Kumar et al.[23] carried out a comprehensive statistical analysis of five main odor databases: Flavornet, GoodScents, Leon & Johnson, Sigma–Aldrich, and SuperScent. One problem with these databases is the sparseness of the data distribution, because an odorant molecule can be described by a varying number of ODs, but very few molecules are described by a large number of ODs in the databases. The statistical results of Kumar et al.[23] indicate that the Sigma–Aldrich database possesses both a relatively larger number of odorant molecules and larger average number of ODs per molecule, and thus it leads to the highest average occurrence of ODs. In view of this characteristic, the Flavors and Fragrances database of Sigma–Aldrich (2016), which has been upgraded to 1026 odorant molecules and 160 ODs, was adopted and analyzed in the present study. Detailed information about the 160 ODs is listed in Table S1. Figure 3 summarizes the 20 ODs that occurred most frequently

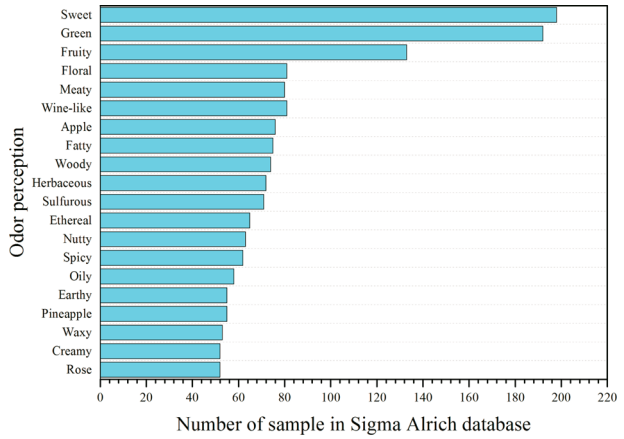


Figure 3. Twenty most frequent ODs in Sigma–Aldrich database. Given the sample size for model calibration, the first 10 odor descriptors were considered in this study.

in the database. The descriptor sweet is represented by approximately 200 odorants, while the descriptor rose is represented by approximately 50 odorants. Considering that a badly calibrated model could result from insufficient sampling, only the top 10 ODs were used to establish our prediction models. The 10 descriptors were sweet, green, fruity, floral, meaty, winelike, apply, fatty, woody, and herbaceous. The minimum number of samples (herbaceous) is over 70. This sample size may help ensure accuracy of the prediction models.

Feature Extraction. As machine learning aims to deal with larger, more complex questions, the extraction of relevant features for data representation data is a critical problem within model calibration.[37] It has been reported that machine-learning algorithms exhibit a decrease in accuracy when the number of variables is significantly higher than an optimal number.[38] Consequently, before model calibration, PCA and BR were employed as unsupervised and supervised methods, separately, to extract features from all the MPs, and their effects were evaluated.

PCA was first performed to remove redundant information (Figure S1). To avoid loss of characteristic information from the original data set, PCs with accumulative contributions of 99.99% were selected. Table 1 lists the number of PCs for 10 ODs. BR was used to find useful features of each OD. By BR, 1006 MPs were labeled as confirmed, tentative, or rejected (Table 1). In this research, MPs labeled confirmed or tentative (BR-CT), and labeled confirmed only (BR-C) were used in further processing. Features selected by BR for the 10 ODs are shown in Figure 4. This illustrates that although an MP may be labeled as confirmed for one OD, the MP could be regarded as a useless feature for other ODs. This indicates that ODs could be used to describe various dimensions for an odorant. It can be interpreted that some MPs are associated with some appointed functional groups of a molecule, and functional groups are related to ODs.

Random Forest Model. To calibrate RF models, two parameters, the number of trees (n_{tree}) and the number of features (m_{try}), need to be optimized. Although adding more trees will not cause overfitting, it will increase the model complexity. Therefore, a sufficient number of trees was needed

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999–12005

Table 1. Data Sets, Division of Samples, Principal Components, and Molecular Parameters[a]

	original data set			SMOTE processed data set		division of samples by KS[b]			MPs labeled by BR method[c]		
odor descriptor	P sample	N sample	N:P	P sample	N sample	train set	test set	no. of PCs	confirmed	tentative	rejected
sweet	198	828	4.18:1	792	891	1262	421	260	180	206	620
green	192	834	4.34:1	768	864	1224	408	267	263	211	532
fruity	133	893	6.71:1	532	598	847	283	238	214	205	587
floral	81	945	11.67:1	324	364	516	172	201	118	136	752
meaty	80	946	11.83:1	320	360	510	170	201	141	208	657
winelike	81	945	11.67:1	324	364	516	172	200	95	126	785
apple	76	950	12.50:1	304	342	484	162	188	133	109	764
fatty	75	951	12.68:1	300	337	477	160	199	116	155	735
woody	74	952	12.86:1	296	333	471	158	189	129	132	745
herbaceous	72	954	13.25:1	288	324	459	153	188	122	177	707

[a]Original and synthetic minority oversampling technique (SMOTE)-processed data sets are described. P (positive sample) indicates the number of samples with the specific OD label. N (negative sample) indicates the number of samples with the nonspecific OD label. [b]Divided by use of the Kennard−Stone (KS) algorithm, [c]Molecular parameters labeled by the Boruta method.

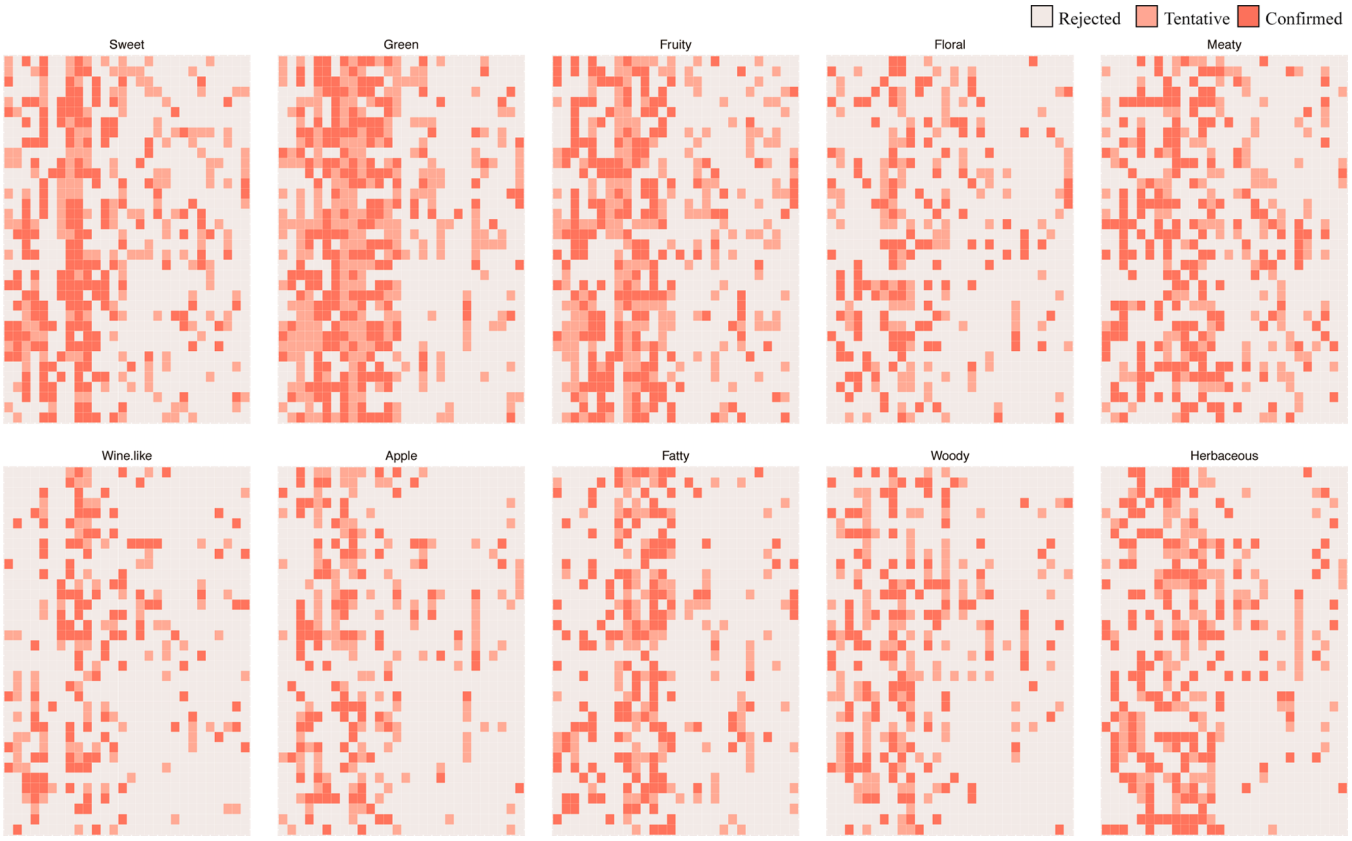


Figure 4. MPs selection based on BR method for the ten ODs. One thousand six MPs were arranged as a matrix (36 × 28). Each grid indicated one MP. MPs labeled as rejected, tentative, or confirmed were colored. Here, the features labeled tentative and confirmed (BR-CT) and only labeled confirmed (BR-C) were used for calibrating models. The number of MPs for each label is listed in Table 1.

to calibrate RF models. With the out-of-bag error and test error taken into consideration, the optimal n_{tree} and m_{try} were determined (Figures S3 and S4). The optimal modeling parameters for the RF models are listed in Table 2. The overall accuracies of the best RF models under all parameters (AP), PCA, BR-CT, and BR-C data sets are shown in Figure 5. In summary, the PCA RF model showed a better average accuracy (92.79% ± 1.63%) than the AP (90.62% ± 1.26%), BR-CT (90.50% ± 1.21%), and BR-C (90.61% ± 1.85%) RF models.

Extreme Learning Machine Model. For ELM models, the parameter that needs to be tuned is the number of hidden layer nodes. In this study, the parameter was obtained by a trial and error method. The range of number of hidden nodes was set from 1 to 800. To avoid the randomness of ELM models, each ELM model was repeated 200 times and the average accuracy was employed to finish the parameter's selection (Figure S5). On the basis of highest average accuracies of calibration set and validation set, the optimal parameter was determined. Selected results are provided in Table 2. The OD identification accuracies for the training and test sets of ELM are shown in

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005

Table 2. Modeling Parameters for Support Vector Machine, Random Forest, and Extreme Learning Machine Model Calibration[a]

	AP					PCA					BR-CT					BR-C				
odor descriptor	c	g	m_{try}	n_{tree}	n_{hidden}	c	g	m_{try}	n_{tree}	n_{hidden}	c	g	m_{try}	n_{tree}	n_{hidden}	c	g	m_{try}	n_{tree}	n_{hidden}
sweet	1.414	0.006	46	187	660	1.414	0.008	160	126	794	1.414	0.011	96	39	575	1.414	0.022	90	52	387
green	2.828	0.004	186	34	575	2.000	0.008	27	97	796	2.000	0.008	124	56	633	5.657	0.022	103	32	440
fruity	2.828	0.008	86	90	537	4.000	0.008	128	91	626	1.414	0.011	99	32	508	256	0.022	54	25	399
floral	1.414	0.006	36	93	320	1.414	0.006	21	133	792	1.000	0.022	34	53	230	1.414	0.011	28	218	204
meaty	1.000	0.001	346	75	326	1.000	0.001	21	59	795	2.000	0.006	109	31	321	2.828	0.004	91	24	271
winelike	4.000	0.006	26	32	347	4.000	0.006	130	170	786	2.000	0.022	11	36	309	1.414	0.088	75	66	322
apple	1.414	0.011	196	54	312	1.414	0.011	128	37	795	1.414	0.022	172	57	244	1.414	0.088	123	36	335
fatty	8.000	0.002	666	27	255	8.000	0.002	129	106	786	4.000	0.006	51	42	220	1.414	0.031	96	34	180
woody	4.000	0.008	46	92	295	2.828	0.008	69	107	797	1.414	0.022	121	61	270	1.000	0.044	79	64	217
herbaceous	4.000	0.008	36	56	328	8.000	0.011	78	87	768	2.000	0.044	29	74	305	2.828	0.063	82	20	724

[a]AP, all parameters; PCA, principal component analysis; BR-CT, confirmed and tentative from Boruta method; BR-C, confirmed only from Boruta method; c, penalizing factor for SVM models; g, RBF kernel parameter for SVM models; m_{try}, number of features for RF models; n_{tree}, number of trees for RF models; n_{hidden}, number of hidden nodes for ELM models.

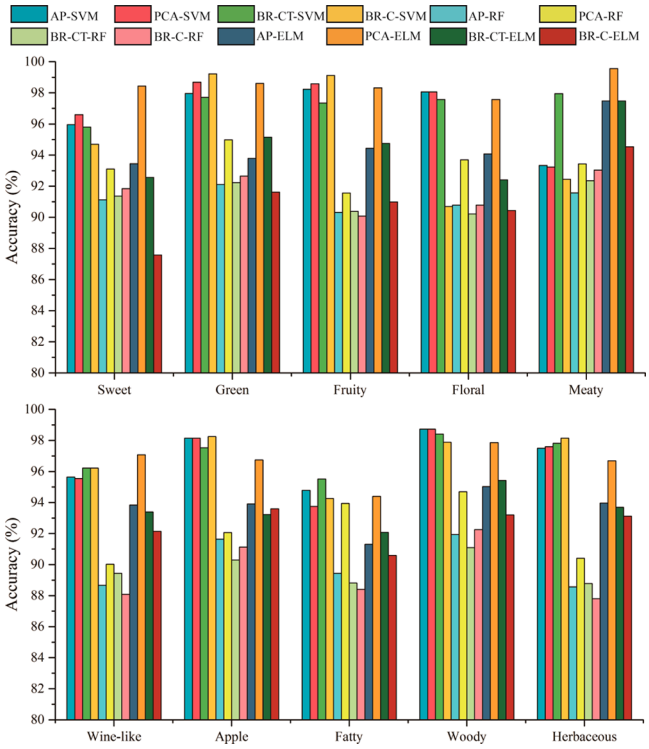


Figure 5. Identification average accuracies of train and test sets for 10 odor descriptors by SVM, RF, and ELM.

Figure 5, which shows that the accuracy of PCA ELM model (97.53% ± 1.35%) is higher than those of AP (94.13% ± 1.44%), BR-CT (94.02% ± 1.59%), and BR-C (91.78% ± 1.91%) ELM models.

Model Comparison. The average accuracies of the training and test sets for ODs are shown in Figure 5. For green, fruity, winelike, apple, and herbaceous identification, BR-C-SVM shows better results than other models. However, PCA ELM did a better job identifying the sweet and meaty ODs. When the tree modeling methods were compared (Figure 6), it was found that ELM had the best identification accuracy (97.53% ± 1.35%), followed by SVM (97.19% ± 0.93%), and RF (92.79%

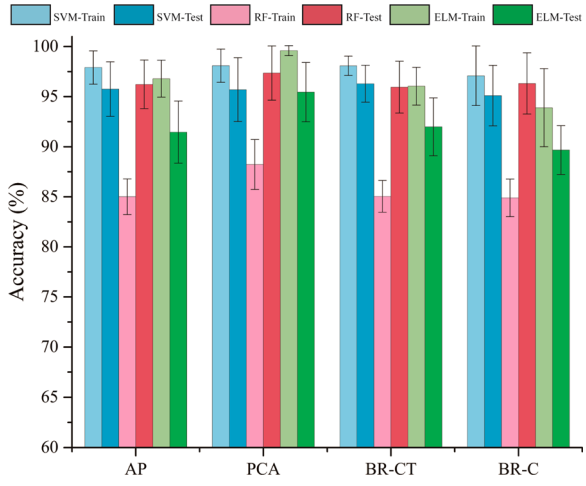


Figure 6. Comparison of average identification accuracies by SVM, RF, and ELM models under all parameters (AP), features extracted by PCA, BR-CT (confirmed or tentative) or BR-C (confirmed only).

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005

± 1.63%). Dealing with large variables slows down machine-learning algorithms and requires more resources.[39] Here, PCA and BR were employed to extract kernel information from a large feature set. The results show that PCA did a better job than Boruta in the RF and ELM models. However, PCA is an unsupervised feature combination method; the PCs are computed on the basis of the original data set. When the amount of input information is considered, BR is more suitable for feature extraction from MPs. It was confirmed that training time increases with the number of features. Here, by the BR-C method, only 15.01% information was extracted instead of all MPs. Therefore, when the accuracies and modeling time are considered comprehensively, it is suggested that SVM combined with features extracted by BR-C, whose average accuracy was higher than 96.10% ± 2.8%, is the optimal model in identifying perceptual descriptors based on MPs. Besides, the recall, precision, and F1 score of BR-C SVM were 94.83% ± 5.61%, 86.88% ± 3.04%, and 95.74% ± 3.52%, respectively, which yields a model with acceptable generalization ability to predict odor descriptors based on physicochemical parameters.

Model Validation. Thirty primary VOCs, identified from Golden Delicious apples analyzed by GC/MS, and their ODs, from databases and predicted by the models in this study, are summarized in Table 3, which indicates that 70% (21/30) of compounds were predicted accurately. The other seven compounds were shown to be unpredictable, which can be explained by the insufficient number of OD models calibrated in presented research. Some ODs, such as peanut and balsamic, were not considered in the present study because of their smaller samples. Although 70% would not be enough to use the model instead of panelists for GC/O, the ODs predicted by models can apply references for the panelists to enhance their work efficiency. Additionally, the predicted accuracy would be increased by consideration of more odorant samples and establishment of enough OD models.

Possibility of Machine-Learning-Based Gas Chromatography/Olfactometry. This paper reported a proof-of-concept study aimed at testing the feasibility of machine-learning-based GC/O. In this study, 10 ODs were tested due to their relatively larger sample size and higher occurrence frequency in the selected database. A SVM model combined with feature extraction by BR-C showed high accuracy in the prediction of these ODs from the MPs of odorant molecules. Although 10 descriptors are obviously not enough for a practical application, the results of this study demonstrate the possibility that a machine-learning approach can be used to obtain sensory information on GC effluents. Additionally, more models for ODs would be expected if sufficient samples can be acquired. About 3000 odorants with odor types were reported in existing databases include Flavornet, GoodScents and SuperScent. In future work, in order to extend the prediction models to more ODs, more odorants with OD information would be collected and a summarized odor database would be established. In fact, the number of prediction models actually needed in a GC/O system may be not as many as expected. A recent study suggested that the dimensionality of odor percepts may be around 20 or less, although the human nose has 400 olfactory receptors.[40] This may mean that 20 or fewer descriptors are enough for their application in GC/O. As shown in Table 3, multiple ODs can be predicted for one compound. These ODs can be used as reference for panelists to obtain a relatively credible odor evaluation, which can enhance their work efficiency and accuracy. In future, a reliable enough model system would be established to rely on adequate samples instead of panelists for odor type evaluation. In conclusion, the prediction model combined with descriptor indexing may be a good candidate to replace the human panelist in GC/O.

Table 3. Model Validation by Golden Delicious Apple Sample[a]

no.	volatile organic compound	odor descriptor from database[b]	predicted odor descriptor
1	2-propanol	alcohol, butter	
2	**1-propanol**	alcohol, **apple**, musty, earthy, peanut, pear, sweet	**apple**
3	**1-butanal**	apple, chocolate, creamy, **green**, meaty, ethereal	**green**, fruity
4	ethyl acetate	solventlike, fruity, anise, ethereal, pineapple	
5	**2-methyl-1-propanol**	**fruity**, whiskey, **winelike**, solventlike	**fruity**, **winelike**
6	1-butanol	banana, vanilla, fruity	
7	**propyl acetate**	**fruity**, floral	**fruity**
8	2-methyl-1-butanol	onion, malty	
9	1-pentanol	sweet, vanilla, balsamic	
10	**isobutyl acetate**	**apple**, banana, ethereal, pear, pineapple	**apple**
11	**1-hexanal**	**fatty**, **green**	**green**, **fatty**
12	**butyl acetate**	banana, **green**, sweet	**green**
13	**(*E*)-2-hexen-1-al**	almond, **apple**, **green**, vegetable	**green**, **apple**, fatty
14	**1-hexanol**	**green**, **herbaceous**, **woody**	**green**, fatty, **woody**, **herbaceous**
15	**2-methyl-1-butyl acetate**	banana, peanut, fruity, applelike	
16	butyl propanoate	banana, ethereal	apple
17	**amyl acetate**	**fruity**, banana, earthy, ethereal	**fruity**, apple
18	**(*E*)-2-hepten-1-al**	**fruity**, rose, **fatty**, almondlike	green, **fruity**, apple, **fatty**
19	6-methyl-5-hexen-2-one	fruity, citruslike, strawberry	
20	**butyl butanoate**	**apple**, banana, berry, peach, pear	**apple**
21	**hexyl acetate**	**apple**, banana, cherry	**apple**, fatty
22	2-ethyl-1-hexanol	oily, rose, sweet	woody, herbaceous
23	**butyl 2-methyl butanoate**	**apple**, chocolate	**apple**
24	**1-octanol**	**fatty**, citrus, waxy, **woody**	**fatty**, **woody**
25	**1-nonanal**	apple, coconut, **fatty**, fishy	**fatty**
26	**hexyl butanoate**	green, **fruity**, **apple**, waxy	**fruity**, winelike, **apple**, fatty
27	***p*-allylanisole**	alcohol, **green**, minty, **sweet**, vanilla	**sweet**, **green**, floral
28	**hexyl 2-methyl butanoate**	**green**, **fruity**, **apple**, grapefruitlike	**green**, **fruity**, **apple**, herbaceous
29	**hexyl hexanoate**	**green**, vegetable, **fruity**, apple, cucumberlike	**green**, **fruity**, fatty
30	(*E*,*E*)-α-farnesene	green, herbaceous	

[a]Boldface type indicates correctly predicted ODs. [b]The odor databases included Flavornet, Sigma−Aldrich, GoodScents, and SuperScent.

CONCLUSIONS

The contribution of this study is to present an approach to predict odor perceptions on the basis of physicochemical descriptors. After processing by the SMOTE and KS methods for balancing data set and subset partitioning, two feature

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005

extraction methods (PCA and BR) were used to extract kernel information from 1006 MPs. Three machine-learning approaches (SVM, RF, and ELM) were employed to establish odor descriptor classifier models. The results showed that models calibrated by SVM presented better accuracies than others. Although the accuracy of the BR-C SVM model is lower than those of AP, PCA, and BR-CT SVM models, when the complexities of the models are considered, BR-C SVM would be the optimal model in this study. Therefore, BR-C SVM has good potential in predicting odor perceptions rapidly and precisely. This study demonstrated that MPs associated with machine-learning models can be adopted for odor perceptual senses identification. The research is expected to offer a novel approach for developing machine-learning-based GC/O.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b02389.

Five figures showing accumulative contribution rates of the first 300 PCs, grid search process of penalty factor and RBF parameter, impact of number of features and number of trees on misclassification error for RF models, and accuracies under different numbers of hidden nodes for ELM models; two tables listing odor perceptions from Sigma−Aldrich database and comparison of results for SVM, RF and ELM models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Telephone +81 092-802-3629; fax +81 092-802-3629; e-mail hayashi@ed.kyushu-u.ac.jp.

ORCID

Liang Shang: 0000-0001-8369-3049

Kenshi Hayashi: 0000-0001-8679-4953

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by a grant from the China Scholarship Council (CSC) and JSPS KAKENHI Grant 15H01713.

■ REFERENCES

(1) Bartsch, J.; Uhde, E.; Salthammer, T. *Anal. Chim. Acta* **2016**, *904*, 98−106.

(2) Brattoli, M.; Cisternino, E.; Dambruoso, P. R.; de Gennaro, G.; Giungato, P.; Mazzone, A.; Palmisani, J.; Tutino, M. *Sensors* **2013**, *13*, 16759−16800.

(3) Zellner, B. D.; Dugo, P.; Dugo, G.; Mondello, L. *J. Chromatogr.* **2008**, *1186*, 123−143.

(4) Casilli, A.; Decorzant, E.; Jaquier, A.; Delort, E. *J. Chromatogr.* **2014**, *1373*, 169−178.

(5) Erten, E. S.; Cadwallader, K. R. *Food Chem.* **2017**, *217*, 244−253.

(6) Acree, T. E. *Anal. Chem.* **1997**, *69*, 170A−175A.

(7) Chin, S. T.; Eyres, G. T.; Marriott, P. J. *Anal. Chem.* **2012**, *84*, 9154−9162.

(8) Delahunty, C. M.; Eyres, G.; Dufour, J. P. *J. Sep. Sci.* **2006**, *29*, 2107−2125.

(9) Guth, H. *J. Agric. Food Chem.* **1997**, *45*, 3022−3026.

(10) Pollien, P.; Fay, L. B.; Baumgartner, M.; Chaintreau, A. *Anal. Chem.* **1999**, *71*, 5391−5397.

(11) van Ruth, S. M. *Biomol. Eng.* **2001**, *17*, 121−128.

(12) Korichi, M.; Gerbaud, V.; Floquet, P.; Meniai, A. H.; Nacef, S.; Joulia, X. In *16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on Process Systems Engineering*, Marquardt, W.; Pantelides, C., Eds.; Computer Aided Chemical Engineering, Vol. *21*; Elsevier: 2006; pp 895−900.

(13) Malnic, B.; Hirono, J.; Sato, T.; Buck, L. B. *Cell* **1999**, *96*, 713−723.

(14) Luo, S. X.; Axel, R.; Abbott, L. F. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 10713−10718.

(15) Johnson, B. A.; Xu, Z.; Ali, S. S.; Leon, M. *J. Comp. Neurol.* **2009**, *514*, 658−673.

(16) Mori, K.; Takahashi, Y. K.; Igarashi, K. M.; Yamaguchi, M. *Physiol. Rev.* **2006**, *86*, 409−433.

(17) Teixeira, M. A.; Barrault, L.; Rodriguez, O.; Carvalho, C. C.; Rodrigues, A. E. *Ind. Eng. Chem. Res.* **2014**, *53*, 8890−8912.

(18) Arzi, A.; Sobel, N. *Trends Cognit. Sci.* **2011**, *15*, 537−545.

(19) Kim, S. J.; Shin, D. H. In *HCI in Business, Government, and Organizations: eCommerce and Innovation, Pt I*, Nah, F. F. H., Tan, C. H., Eds.; Springer: 2016; pp 406−416; DOI: 10.1007/978-3-319-39396-4_37.

(20) Mamlouk, A. M.; Martinetz, T. *Neurocomputing* **2004**, *58-60*, 1019−1025.

(21) Pintore, M.; Wechman, C.; Sicard, G.; Chastrette, M.; Amaury, N.; Chretien, J. R. *J. Chem. Inf. Model.* **2006**, *46*, 32−38.

(22) Haddad, R.; Khan, R.; Takahashi, Y. K.; Mori, K.; Harel, D.; Sobel, N. *Nat. Methods* **2008**, *5*, 425−429.

(23) Kumar, R.; Kaur, R.; Auffarth, B.; Bhondekar, A. P. *PLoS One* **2015**, *10*, e0141263.

(24) Keller, A.; Gerkin, R. C.; Guan, Y. F.; Dhurandhar, A.; Turu, G.; Szalai, B.; Mainland, J. D.; Ihara, Y.; Yu, C. W.; Wolfinger, R.; Vens, C.; Schietgat, L.; De Grave, K.; Norel, R.; Stolovitzky, G.; Cecchi, G. A.; Vosshall, L. B.; Meyer, P. *Science* **2017**, *355*, 820−826.

(25) Teixeira, M. A.; Rodriguez, O.; Rodrigues, A. E.; Selway, R. L.; Riveroll, M.; Chieffi, A. *Ind. Eng. Chem. Res.* **2013**, *52*, 963−971.

(26) Teixeira, M. A.; Rodriguez, O.; Rodrigues, A. E. *Ind. Eng. Chem. Res.* **2010**, *49*, 11764−11777.

(27) Ferreira, V.; Ortin, N.; Escudero, A.; Lopez, R.; Cacho, J. *J. Agric. Food Chem.* **2002**, *50*, 4048−4054.

(28) Sigma-Aldrich. Flavors and Fragrances. Natural and Food Grade Ingredients. 2016, http://www.sigmaaldrich.com/industries/flavors-and-fragrances.html.

(29) Li, J. Y.; Fong, S. M.; Sung, Y. S.; Cho, K. G.; Wong, R.; Wong, K. K. L. *BioData Min.* **2016**, *9*, No. 37, DOI: 10.1186/s13040-016-0117-1.

(30) Kaneko, H.; Funatsu, K. *Chemom. Intell. Lab. Syst.* **2016**, *153*, 75−81.

(31) Li, X. H.; Kong, W.; Shi, W. M.; Shen, Q. *Chemom. Intell. Lab. Syst.* **2016**, *155*, 145−150.

(32) Agjee, N. H.; Ismail, R.; Mutanga, O. *J. Appl. Remote Sens* **2016**, *10*, No. 042002.

(33) Poona, N. K.; van Niekerk, A.; Nadel, R. L.; Ismail, R. *Appl. Spectrosc.* **2016**, *70*, 322−333.

(34) Ren, J. C. *Knowl. Based Syst.* **2012**, *26*, 144−153.

(35) Arvisenet, G.; Billy, L.; Poinot, P.; Vigneau, E.; Bertrand, D.; Prost, C. *J. Agric. Food Chem.* **2008**, *56*, 3245−3253.

(36) Kaeppler, K.; Mueller, F. *Chem. Senses* **2013**, *38*, 189−209.

(37) Roweis, S. T.; Saul, L. K. *Science* **2000**, *290*, 2323−2326.

(38) Janus, M.; Morawski, A. W. *Appl. Catal., B* **2007**, *75*, 118−123.

(39) Cheng, X. P.; Cai, H. M.; Zhang, Y.; Xu, B.; Su, W. F. *BMC Bioinf.* **2015**, *16*, No. 219.

(40) Meister, M. *eLife* **2015**, *4*, No. e07865.

DOI: 10.1021/acs.analchem.7b02389
Anal. Chem. 2017, 89, 11999−12005